Markov Decision Processes

$p(s', r|s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$ — one step dynamic

$p(s'|s, a) = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a) = \sum_{r \in \mathcal{R}} p(s', r|s, a)$ — next state distribution

$r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)$ — expected reward for taking action a in the state s

$\mathbb{E}[G_t] = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}]$ — expected return

Value Functions

$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s]$

$Q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a]$

Bellman Expectation Equations

expected return from a given state s following the policy π

$V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s]$

$\quad = \sum_{a \in \mathcal{A}} \pi(a|s) \; [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_\pi(s')]$

$V_\pi = \pi(R + \gamma P V_\pi)$

$\boxed{V_\pi = (I - \gamma \pi P)^{-1} \pi R}$

expected return from a given state s when action a is computed and then π is followed

$Q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s, A_t = a]$

$\quad = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_\pi(s')$

$\quad = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \sum_{a \in \mathcal{A}} \pi(a'|s') Q_\pi(s', a')$

$Q_\pi = R + \gamma P\pi Q_\pi$

$\boxed{Q_\pi = (I - \gamma P\pi)^{-1} R}$

Optimality

$V^*(s) = \max_\pi V_\pi(s) \qquad \forall s \in \mathcal{S}$

$Q^*(s, a) = \max_\pi Q_\pi(s, a) \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$

Bellman Optimality Equations

$V^*(s) = \sum_{a \in \mathcal{A}} \pi^*(a|s) \; [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V^*(s')]$

$\quad = \max_a [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V^*(s')]$

$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \sum_{a \in \mathcal{A}} \pi^*(a'|s') Q^*(s', a')$

$\quad = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \max_{a'} Q^*(s', a')$

Optimal policy

$\pi^*(s) = \arg\max_a [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V^*(s')] = \arg\max_a Q^*(s, a)$

Dynamic Programming

Policy Iteration — starts from a random policy

Evaluation: $V_{k+1}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \; [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_k(s')] \qquad \forall s \in \mathcal{S}$

Improvement: $\pi'(s) = \arg\max_a [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_\pi(s')] = \arg\max_a Q_\pi(s, a) \qquad \forall s \in \mathcal{S}$

Value Iteration — interleave partial evaluation and partial improvement

$V_{k+1}(s) \leftarrow \max_a [r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_k(s')] \qquad \forall s \in \mathcal{S}$

(*)

Reinforcement Learning

Prediction

Monte Carlo (first/every visit): $V(S_t) \leftarrow average[G_t | S_t]$

Temporal Difference (TD(0)): $V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$

combines sampling (from MC) and bootstrapping (from DP): learn/update the value function in s based on the value function of the states next to s

Control

Monte Carlo: $Q(S_t, A_t) \leftarrow average[G_t | S_t, A_t]$

SARSA: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$

Q-Learning: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t))$

Improvement: ϵ-greedy algorithm

$$\pi(a|s) = \begin{cases} \arg\max_a Q(s,a) & \text{with probability } 1 - \varepsilon + \frac{\varepsilon}{|A(s)|} \\ \text{not } (\uparrow) \text{ (something random)} & \text{〃} \quad \frac{\varepsilon}{|A(s)|} \end{cases}$$

Off-policy learning

- learning π
- behavior policy b

$$V_\pi(s) \approx \frac{\sum_n \rho_n \text{Return}_n}{N}$$

$$\rho_n = \frac{\mathbb{P}(\text{trajectory under } \pi)}{\mathbb{P}(\text{trajectory under } b)}$$

Multi-Armed Bandit

- **Stochastic MAB**

  $L_T = T \cdot R^* - \mathbb{E}[\sum_{t=1}^{T} R(a_{i_t})] = \sum_{a \in \mathcal{A}} \mathbb{E}[N_T(a_i)]\Delta_i$

  Lower Bound: $\lim_{T\to\infty} L_T \geq \log T \sum_{a_i|\Delta_i>0} \frac{\Delta_i}{KL(R(a_i),R(a^*))}$

similar upper bounds

Upper Confidence Bound 1 (UCB1) – *frequentist approach*

$\propto$ inversely to the # pulls

For each time step t : compute : $\hat{R}_t(a_i) = \frac{\sum_{i=1}^{t} r_{i,t}\mathbb{I}_{a_i=a_{i_t}}}{N_t(a_i)}$ $\forall a_i$

$B_t(a_i) = \sqrt{\frac{2\log t}{N_t(a_i)}}$ $\forall a_i$

play arm : $a_{i_t} = \arg\max_{a_i \in \mathcal{A}}(\hat{R}_t(a_i) + B_t(a_i))$

Upper Bound: $L_T \leq 8 \log T \sum_{i|\Delta_i>0} \frac{1}{\Delta_i} + (1+\frac{\pi^2}{3}) \sum_{i|\Delta_i>0} \Delta_i$

Thomson Sampling – *bayesian approach*

Consider a bayesian prior for each arm $f_1, .., f_N$ as a starting point. At each round t we sample from each one of the distributions, obtaining $\hat{r}_1, .., \hat{r}_N$. We pull the arm a_{i_t} with the highest sampled value $i_t = \arg\max_i \hat{r}_i$. Then we update the prior incorporating the new information.

In the case of Thomson sampling for Bernoulli rewards we use as prior conjugate distributions the $Beta(\alpha,\beta)$ and the *Bernoulli*. We start from all equal priors for all arms: $f_i(0) = Beta(\alpha_0 = 1, \beta_0 = 1) = \mathcal{U}([0,1])$. Then, when we pull an arm i, if we obtain a success we update $f_i(t+1) = Beta(\alpha_t + 1, \beta_t)$, if instead we obtain a failure we update $f_i(t+1) = Beta(\alpha_t, \beta_t + 1)$.

Upper Bound: $L_T \leq O(\sum_{i|\Delta_i>0} \frac{\Delta_i}{KL(\mathcal{R}(a_i),\mathcal{R}(a^*))}(\log T + \log\log T))$

- **Adversarial MAB**

  $L_T = \max_i \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{i_t,t}$

  Lower Bound: $\inf \sup \mathbb{E}[L_T] \geq \frac{1}{20}\sqrt{T \cdot N}$

EXP3

$\pi_t(a_i) = (1-\beta)\frac{w_t(a_i)}{\sum_j w_t(a_j)} + \frac{\beta}{N}$ where: $w_{t+1}(a_i) = \begin{cases} w_t(a_i)e^{\eta \frac{r_{i,t}}{\pi_t(a_i)}} & \text{if } a_i \text{ has been pulled} \\ w_t(a_i) & \text{if else} \end{cases}$

Upper Bound: $\mathbb{E}[L_T] \leq O(\sqrt{T \cdot N \log N})$ with: $\beta = \eta = \sqrt{\frac{N\log N}{(\epsilon-1)T}}$